

〈빅데이터분석실습 과제01〉

문제 : “daily_88101_2016.csv” 는 미국에서 측정된 2016년 일별(daily) 공기 중의 오염도에 대한 데이터이다.
각 county(County.Code)별로 AQI의 월별 평균을 구하라.

- Data.Local 변수에서 측정된 월을 추출한다.
month = months(as.POSIXlt(Date.Local))
- group_by()함수를 County.Code와 추출된 month변수에 적용하여 층(Strata)을 구성하고, summary() 함수를 이용하여 월 평균을 구한다.

데이터 구조

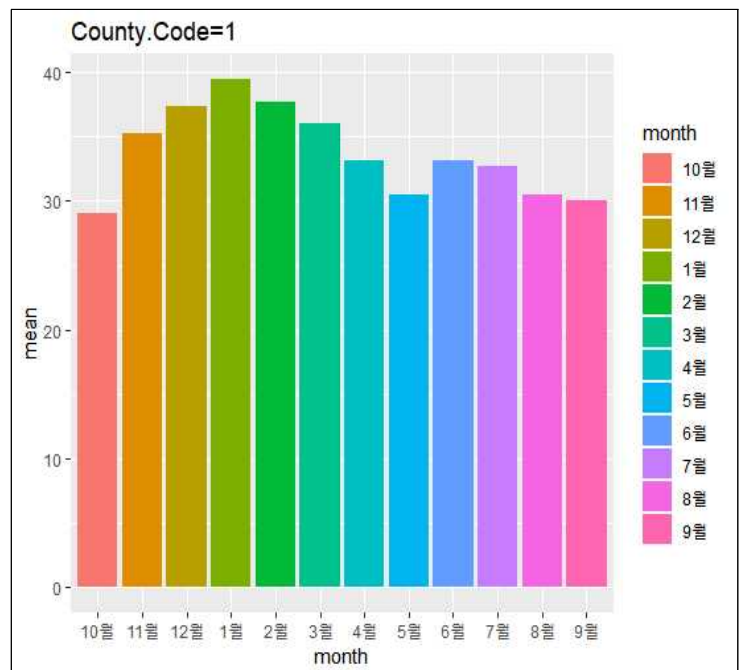
- 371,132 obs. of 29 variables

	State.Code	County.Code	Site.Num	...	CBSA.Name	Date.of.Last.Change
1	1	3	10	...	Daphne-Fairhope-Foley, AL	2015-05-01
2	1	3	10	...	Daphne-Fairhope-Foley, AL	2015-05-01
3	1	3	10	...	Daphne-Fairhope-Foley, AL	2015-05-01
...
371,132	78	10	12			2016-11-21

County.Code에 따른 month 별 AQI 평균 값

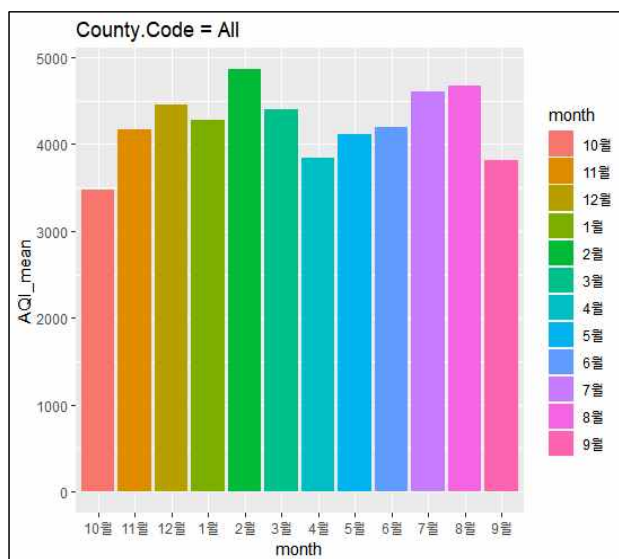
- County.Code는 이산형 변수로 1,2,3,...,810 까지의 자연수로 표현되어 있다.
- month는 Date.Local에서 추출한 변수로 문자형 변수이며 “1월”부터 “12월” 까지 표현되어 있다.

County.Code	month	AQI_Mean
1	10월	29.077905
1	12월	35.208666
1	12월	37.344411
1	1월	39.469018
1	2월	37.672043
1	3월	36.055609
1	4월	33.142703
1	5월	30.466019
1	6월	33.171397
1	7월	32.659066
1	8월	30.538462
1	9월	30.021226
3	10월	24.815904
3	11월	33.334816
... 1446 obs		



전체 month 별 AQI 평균 값

월	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
평균	39.1	39.8	34.7	30.2	31.4	32.8	36.5	36.1	31.3	27.9	34.7	36.0



County.code와 month와의 관계

County.Code	month											
	10월	11월	12월	1월	2월	3월	4월	5월	6월	7월	8월	9월
1	1430	1597	1804	1905	1673	1895	1688	1662	1644	1658	1492	1563
3	1534	1507	1619	1557	1391	1514	1515	1556	1487	1595	1601	1533
5	1072	972	1087	832	871	976	901	902	834	965	1061	1055
7	720	781	847	666	563	664	709	672	694	694	714	635
9	732	723	757	773	736	757	769	864	780	848	812	801
10	4	5	6	5	0	0	3	5	5	5	5	3
11	427	436	377	440	410	411	436	437	388	414	471	432
12	62	60	62	62	56	62	60	62	60	32	61	59
13	1097	1079	1134	982	908	1002	1034	1042	1060	1108	1098	1054
15	286	275	270	292	258	293	266	367	308	346	335	335
17	349	309	256	261	262	290	277	278	291	298	296	276
19	600	605	642	468	486	574	560	566	568	589	567	564
20	124	116	124	108	111	124	112	140	120	110	122	117
21	600	561	605	548	493	531	544	556	524	560	586	584
23	231	258	291	192	211	279	274	271	277	281	285	228

R 코드

패키지 호출

```
library(tidyverse)
```

데이터 호출

```
data <- read.csv("daily_88101_2014.csv")
```

데이터 구조 파악

```
str(data)
```

```
summary(data)
```

데이터 미리보기

```
head(data)
```

```
tail(data)
```

Date.Local 변수에서 월을 추출하여 새로운 변수로 지정하기

```
data <- data %>% mutate(month=months(as.POSIXlt(data$Date.Local)) )
```

group_by() 함수를 이용하여 층(strata) 구성하여 AQI 평균 구하기

```
data %>% group_by(County.Code, month) %>% summarise(AQI_mean=mean(AQI,na.rm=TRUE))
```

전체 월별 평균

```
data %>% group_by(month) %>%  
  summarise(AQI_mean=mean(AQI,na.rm=TRUE))
```

County.Code와 month와의 관계

```
with(data,table(County.Code,month))
```

전체 월별 평균 막대그래프

```
data %>% group_by(County.Code, month) %>%  
  summarise(AQI_mean=mean(AQI,na.rm=TRUE)) %>%  
  ggplot(aes(x=month,y=AQI_mean,fill=month)) +  
  geom_bar(stat="identity") +  
  labs(title="County.Code = All")
```

County.Code=1 경우의 월별 평균 막대그래프

```
data %>% filter(County.Code==1) %>%  
  group_by(County.Code, month) %>%  
  summarise(AQI_mean=mean(AQI,na.rm=TRUE)) %>%  
  ggplot(aes(x=month,y=AQI_mean,fill=month)) +  
  geom_bar(stat="identity") +  
  labs(title="County.Code=1")
```